METODE BAYESIAN
PADA MODEL REGRESI LINEAR

oleh

TRIWIK JATU P

NIM. M0102008

SKRIPSI

ditulis dan diajukan untuk memenuhi sebagian persyaratan

memperoleh gelar Sarjana Sains Matematika

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA dan ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2007

SKRIPSI

METODE BAYESIAN PADA MODEL REGRESI LINEAR

yang disiapkan dan disusun oleh

TRIWIK JATU PARMANINGSIH

M0102008

dibimbing oleh

Pembimbing I,	Pembimbing II,
Hasih Pratiwi, M.Si	Dr. Sutanto, DEA
NIP. 132 143 817	NIP. 132 149 079

telah dipertahankan di depan Dewan Penguji

pada hari Jumat, tanggal 5 Januari 2007

dan dinyatakan telah memenuhi syarat.

Anggota Tim Penguji	Tanda Tangan
1. Dra. Yuliana Susanti, M. Si	1. .....
NIP. 131 695 845	
2. Dra. Respatiwulan, M. Si	2. .....
NIP. 132 046 022	
3. Drs. Sutrima, M. Si	3. .....
NIP. 132 046 018	

Disahkan oleh

Fakultas Matematika dan Ilmu Pengetahuan Alam

Dekan,	Ketua Jurusan Matematika,
Drs. Marsusi, M.S	Drs. Kartiko, M.Si
NIP. 130 906 776	NIP. 131 569 203

ABSTRAK

Triwik Jatu Parmaningsih, 2007. METODE BAYESIAN PADA MODEL REGRESI LINEAR. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret.

Analisis regresi merupakan teknik statistik yang digunakan untuk menyelidiki hubungan antar variabel. Metode yang biasa digunakan untuk mengestimasi parameter regresi adalah metode kuadrat terkecil. Ketika melakukan informasi tentang parameter populasi, terkadang diperoleh informasi tambahan mengenai parameter populasi yang berasal dari data sebelumnya. Jika informasi tersebut ingin dimasukkan dalam analisis data, maka estimasi dengan metode kuadrat terkecil tidak dapat digunakan. Oleh karena itu, diperlukan metode Bayesian untuk menyelesaikan permasalahan tersebut. Tujuan dari penulisan skripsi ini adalah menentukan estimasi parameter regresi dengan metode Bayesian, menguji signifikansi parameter regresi serta membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

Metode yang digunakan dalam penulisan skripsi ini adalah studi literatur. Langkah yang dilakukan adalah menjelaskan metode Bayesian, mengestimasi parameter regresi melalui harga harapan distribusi *posterior*, menentukan probabilitas *posterior* untuk menguji signifikansi parameter regresi serta melakukan analisis eror untuk membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

Berdasar pembahasan diperoleh kesimpulan bahwa estimasi parameter regresi linear dengan metode Bayesian sama dengan estimasi kuadrat terkecil jika informasi *prior* tidak diketahui, sedangkan jika informasi *prior* diketahui maka estimasi Bayes merupakan harga harapan distribusi *posterior*. Uji hipotesis pada metode Bayesian dilakukan dengan menghitung probabilitas *posterior* $Prob(H_0|\boldsymbol{y})$. Jika $Prob(H_0|\boldsymbol{y}) < Prob(H_0) = \frac{1}{2}$ maka H_0 ditolak.

Kata kunci: *distribusi prior, distribusi posterior, estimator Bayes.*

ABSTRACT

Triwik Jatu Parmaningsih, 2007. BAYESIAN METHOD ON LINEAR REGRESSION MODEL. Faculty of Mathematics and Natural Sciences, Sebelas Maret University.

Regression analysis is a statistical technique for investigating the relationship between variables. The least square method is commonly used to estimate regression parameters. Sometimes, when drawing an inference from population we get an additional information about parameters available from a previous study. If the previous information will be incorporated in data analysis, then the estimation by the least square method can not be used. Therefore, we need Bayesian method to solve such problem. The aims of this research are to estimate regression parameters, to test significance of regression parameters by Bayesian method and to compare the estimation of regression model based on noninformative prior distribution and conjugate prior distribution in a case.

The method used in this research is literary study. The steps are to explain Bayesian method, to estimate regression parameters by the mean of posterior distribution, to determine posterior probability for testing significance of regression parameters and to analyze the error to compare the estimation of regression model based on noninformative prior distribution and conjugate prior distribution in a case.

This research indicates that Bayes estimation turns out to be exactly the same as the least square estimate if prior information is not known, whereas if prior information is known then Bayes estimation is taken to be the expectation of posterior distribution. The hypothesis test in Bayesian method is based on posterior probability calculation $Prob(H_0|\boldsymbol{y})$. The Bayesian test rejects the null hypotesis if $Prob(H_0|\boldsymbol{y}) < Prob(H_0) = \frac{1}{2}$.

***Key words**: prior distribution, posterior distribution, Bayes estimator.*

PERSEMBAHAN

Kupersembahkan tulisan sederhana ini untuk
Bapak dan Ibu, atas segala pengorbanan, doa dan cinta yang tulus.
Mbak Enggar dan mas Sas, yang selalu mendukungku.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah Subhanahu wa Ta'ala, karena atas limpahan rahmat dan karunia-Nyalah skripsi ini dapat diselesaikan.

Ucapan terima kasih tidak lupa penulis sampaikan kepada

1. Hasih Pratiwi, M.Si selaku dosen pembimbing I dan Dr. Sutanto, DEA selaku dosen pembimbing II atas bimbingan, bantuan dan motivasinya.
2. Dra. Yuliana Susanti, M. Si selaku Pembimbing Akademik atas bantuan dan motivasinya.
3. Teman-teman angkatan 2002 yang telah memberi bantuan dan motivasi.
4. Intan, Rettob, Fitria, Hesti, Yuyun terimakasih atas persahabatan yang indah yang telah diberikan.
5. Martopo yang telah memberi bantuan dalam pengoperasian LaTeX.
6. Mbak Imah yang telah meluangkan waktu untuk memberi solusi.
7. Semua pihak yang telah membantu, memudahkan dan memperlancar penulisan skripsi ini.

Disadari bahwa tiada satu pun manusia yang sempurna, maka segala kritik dan saran yang bersifat membangun sangat diharapkan demi perbaikan skripsi ini.

Surakarta, Januari 2007

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Analisis regresi merupakan teknik statistik yang digunakan untuk menyelidiki hubungan antar variabel. Penerapan regresi dapat ditemukan pada beberapa bidang ilmu pengetahuan. Sebagai contoh, pada bidang kedokteran ingin diketahui apakah tekanan darah seseorang bergantung pada umur, berat badan dan tinggi badan. Pada bidang industri ingin diketahui apakah bahan kimia yang dihasilkan bergantung pada waktu reaksi dan suhu reaksi (Birkes dan Dodge [3]).

Untuk mengetahui ada tidaknya hubungan antar variabel, pada umumnya digunakan suatu model yang merepresentasikan hubungan fungsional antar variabel. Regresi merupakan salah satu alat untuk membentuk model tersebut. Variabel yang terkait dalam regresi adalah variabel bebas X dan variabel tak bebas Y. Persamaan $Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$, dengan p menunjukkan banyak variabel bebas, merupakan model regresi dimana f disebut fungsi regresi dan ε adalah sesatan random. Model tersebut disebut model stokastik karena mengandung faktor acak ε. Contoh model regresi linear adalah $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$. Kata linear digunakan untuk menunjukkan bahwa model tersebut linear dalam parameter $\beta_0, \beta_1, \ldots, \beta_p$. Koefisien $\beta_0, \beta_1, \ldots, \beta_p$ merupakan parameter yang tidak diketahui dari model dan disebut dengan koefisien regresi. Sesatan random ε merupakan variabel random dengan rata-rata 0.

Tujuan utama dari analisis regresi adalah mengestimasi parameter yang tidak diketahui dalam model regresi (Montgomery dan Peck [7]). Ada beberapa metode untuk mengestimasi parameter dan menguji parameter regresi, salah satu metode yang sering digunakan adalah metode kuadrat terkecil. Pada metode

tersebut, data merupakan satu-satunya sumber informasi yang akan dibawa ke dalam perhitungan untuk mengestimasi maupun menguji suatu model. Namun pada prakteknya, ketika melakukan inferensi tentang parameter populasi kadang-kadang diperoleh informasi tambahan mengenai parameter populasi, dimana informasi tersebut berasal dari data sebelumnya. Jika informasi tersebut ingin dimasukkan dalam analisis data, maka estimasi dengan metode kuadrat terkecil tidak dapat digunakan. Oleh karena itu, diperlukan metode Bayesian untuk menyelesaikan permasalahan tersebut. Menurut Birkes dan Dodge [3], estimasi parameter maupun uji hipotesis metode Bayesian dihasilkan dengan mengkombinasikan data sekarang dan informasi dari data sebelumnya atau informasi yang sudah ada.

Secara praktis, analisis dengan metode Bayesian memerlukan distribusi *prior* dan distribusi *posterior*. Ada dua macam distribusi *prior* yaitu distribusi *prior* noninformatif dan distribusi *prior* sekawan. Berawal dari distribusi *prior* akan diperoleh distribusi *posterior* yang akan menentukan estimasi parameter. Skripsi ini memberikan kajian ulang estimasi parameter regresi dengan menggunakan metode Bayesian dan memberikan contoh penerapan pada data sekunder.

1.2 Perumusan Masalah

Dari uraian yang telah diberikan dalam latar belakang masalah, dapat dirumuskan permasalahan sebagai berikut

1. bagaimana menentukan estimasi parameter regresi dengan metode Bayesian,
2. bagaimana menguji signifikansi parameter regresi,
3. bagaimana membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

1.3 Batasan Masalah

Permasalahan yang akan dibahas dalam penulisan ini dibatasi pada model regresi linear dan uji signifikansi parameter regresi berdasar distribusi *prior* noninformatif.

1.4 Tujuan Penulisan

Tujuan dari penulisan skripsi ini adalah

1. dapat menentukan estimasi parameter regresi dengan metode Bayesian,
2. dapat menguji signifikansi parameter regresi,
3. dapat membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

1.5 Manfaat

Dari penulisan ini diharapkan dapat memberikan manfaat yaitu secara teoritis dapat menambah pengetahuan tentang metode estimasi parameter regresi alternatif dengan metode Bayesian, dan secara praktis dapat memberi jalan keluar apabila diketahui informasi awal (*prior*) dari suatu data maka informasi tersebut dapat digunakan untuk mengestimasi parameter regresi.

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Untuk mencapai tujuan penulisan, diperlukan pengertian dan teori-teori yang melandasinya. Pada bab ini diberikan penjelasan tentang matriks dan operasinya, konsep dasar statistika, distribusi normal, distribusi normal multivariat, model regresi linear, metode kuadrat terkecil, uji hipotesis, probabilitas bersyarat dan fungsi likelihood, distribusi *prior*, distribusi *posterior*, teorema Bayes, serta estimator Bayes.

2.1.1 Matriks dan Operasi Matriks

Menurut Anton [1], matriks adalah susunan segi empat siku-siku dari bilangan-bilangan yang secara umum dituliskan sebagai berikut

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

a_{11} sampai a_{mn} disebut entri dari matriks $\boldsymbol{A}$ dan dinyatakan dengan elemen umum a_{ij}, $i = 1, 2, \dots, m$ dan $j = 1, 2, \dots, n$. Matriks yang mempunyai m baris dan n kolom disebut matriks berukuran (berorde) m x n.

Pengertian matriks bujur sangkar, matriks identitas, matriks invers, matriks transpos dan matriks simetri diberikan oleh Johnson dan Wichern [4] pada definisi berikut.

Definisi 2.1.1. *Matriks bujur sangkar adalah matriks yang mempunyai jumlah baris sama dengan jumlah kolom.*

Definisi 2.1.2. *Matriks identitas* n x n *yang dinotasikan dengan* $\boldsymbol{I}$ *atau* $\boldsymbol{I_n}$ *adalah matriks bujur sangkar yang elemen-elemen diagonal utamanya adalah satu dan elemen yang lain adalah 0.*

Jika $\boldsymbol{A}$ adalah matriks bujur sangkar orde n dan $\boldsymbol{I}$ adalah matriks identitas orde n, maka perkaliannya adalah $\boldsymbol{IA} = \boldsymbol{AI} = \boldsymbol{A}$.

Definisi 2.1.3. *Invers dari matriks bujur sangkar* $\boldsymbol{A}$ *adalah suatu matriks yang dinotasikan dengan* $\boldsymbol{A}^{-1}$ *sehingga* $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$.

Definisi 2.1.4. *Jika* $\boldsymbol{A}$ *adalah sembarang matriks* m x n *dengan elemen* a_{ij}, $i = 1, 2, \ldots, m$ *dan* $j = 1, 2, \ldots, n$, *maka transpos* $\boldsymbol{A}$ *dinyatakan dengan* $\boldsymbol{A}'$ *berorde* n x m *dengan elemen* a_{ji}, $j = 1, 2, \ldots, n$ *dan* $i = 1, 2, \ldots, m$.
Jadi, jika

$$\boldsymbol{A}_{m\,\mathbf{x}\,n} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix}$$

maka *transpos* dari $\boldsymbol{A}_{m\,\mathrm{x}\,n}$ adalah

$$\boldsymbol{A}'_{n\,\mathbf{x}\,m} = \begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{1n} & \ldots & a_{nm} \end{pmatrix}$$

Definisi 2.1.5. *Matriks bujur sangkar* $\boldsymbol{A}_{n\,\mathrm{x}\,n}$ *dikatakan simetri jika* $\boldsymbol{A} = \boldsymbol{A}'$. *Jadi,* $\boldsymbol{A}$ *simetri jika* $a_{ij} = a_{ji}$, $i = 1, 2, \ldots, n$ *dan* $j = 1, 2, \ldots, n$.

Menurut Johnson dan Wichern [4], vektor adalah susunan bilangan real $u_1, u_2, \ldots, u_n$ yang secara umum dituliskan sebagai berikut

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \qquad \textit{atau} \qquad \mathbf{u}' = \begin{pmatrix} u_1, & u_2, & \ldots, & u_n \end{pmatrix}.$$

Panjang vektor **u** disebut dengan *norm* vektor **u** dan dinotasikan dengan $\| \mathbf{u} \|$. Berdasar teorema Phytagoras, maka panjang vektor $\mathbf{u} = (u_1, u_2)$ dalam ruang dimensi dua adalah

$$\| \mathbf{u} \| = \sqrt{u_1^2 + u_2^2}.$$

2.1.2 Konsep Dasar Statistika

Pengertian tentang ruang sampel, variabel random, variabel random diskrit, variabel random kontinu, fungsi kepadatan probabilitas, harga harapan, fungsi distribusi kumulatif dan variabel random independen diberikan pada beberapa definisi di bawah ini yang diambil dari Bain dan Engelhardt [2].

Definisi 2.1.6. *Himpunan semua hasil (outcomes) yang mungkin dari suatu eksperimen disebut sebagai ruang sampel dan dinotasikan dengan S.*

Definisi 2.1.7. *Variabel random X adalah suatu fungsi yang memetakan setiap hasil e yang mungkin pada ruang sampel S dengan suatu bilangan real x sedemikian sehingga X(e) = x.*

Ada dua macam variabel random, yaitu variabel random diskrit dan variabel random kontinu.

Definisi 2.1.8. *Jika himpunan semua nilai yang mungkin dari variabel random X terhitung, $x_1, x_2, \ldots, x_n$ atau $x_1, x_2, \ldots$, maka X disebut variabel random diskrit. Fungsi*

$$f(x) = P[X = x] \qquad x = x_1, x_2, \ldots$$

menunjukkan probabilitas untuk setiap nilai x dan disebut dengan fungsi kepadatan probabilitas diskrit.

Definisi 2.1.9. *Fungsi distribusi kumulatif variabel random X didefinisikan dengan*

$$F(x) = P[X \leqslant x].$$

Fungsi $F(x)$ sering disebut fungsi distribusi X dan dinotasikan dengan $F_X(x)$. Untuk variabel random diskrit, fungsi distribusinya adalah

$$F_X(x) = \sum_{x_i \leqslant x} f(x_i).$$

Harga harapan variabel random diskrit diberikan pada definisi berikut.

Definisi 2.1.10. *Jika X adalah variabel random diskrit dengan fungsi kepadatan probabilitas f(x), maka harga harapan X didefinisikan sebagai*

$$E(X) = \sum_{x} x f(x).$$

Definisi 2.1.11. *Variabel random X disebut variabel random kontinu jika terdapat fungsi kepadatan probabilitas f(x) sedemikian sehingga fungsi distribusi kumulatif dapat ditunjukkan sebagai*

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Harga harapan variabel random kontinu diberikan pada definisi berikut.

Definisi 2.1.12. *Jika X adalah variabel random kontinu dengan fungsi kepadatan probabilitas f(x), maka harga harapan X didefinisikan sebagai berikut*

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Definisi 2.1.13. *Variabel random $X_1, X_2, \ldots, X_k$ dikatakan independen jika*

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_k \leq X_k \leq b_k) = \prod_{i=1}^{k} P(a_i \leq X_i \leq b_i), \quad \textit{untuk setiap } \ a_i \leq b_i.$$

2.1.3 Distribusi Normal

Definisi distribusi normal di bawah ini diambil dari Bain dan Engelhardt [2].

Definisi 2.1.14. *Distribusi normal dengan rata-rata μ dan variansi σ^2 yang dinotasikan dengan $N(\mu, \sigma^2)$ mempunyai fungsi kepadatan probabilitas*

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Distribusi normal dengan rata-rata $\mu = 0$ dan variansi $\sigma^2 = 1$ disebut distribusi normal standar.

2.1.4 Distribusi Normal Multivariat

Sebagaimana ditulis oleh Seber [8], berdasar pada fungsi kepadatan probabilitas normal untuk satu variabel, yaitu

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-1/2}\exp[-\frac{1}{2\sigma^2}\,(y-\mu)^2], \qquad -\infty < y < \infty \\ &= (2\pi\upsilon)^{-1/2}\exp[-\frac{1}{2}\,(y-\mu)\upsilon^{-1}(y-\mu)], \qquad \upsilon = \sigma^2 > 0 \end{aligned}$$

maka dapat didefinisikan fungsi kepadatan probabilitas multivariat

$$f(y_1, y_2, \ldots, y_p) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{y}-\mu)^{'}\Sigma^{-1}(\boldsymbol{y}-\mu)] \tag{2.1}$$

dengan $-\infty < y_i < \infty$, $i = 1, 2, \ldots, p$. Vektor $\mu_{p\,x\,1}$ merupakan harga harapan vektor random $\boldsymbol{Y}$ dan $\Sigma_{p\,x\,p}$ merupakan matriks variansi kovariansi.

Definisi 2.1.15. *Jika vektor variabel random* $\boldsymbol{Y}$ *mempunyai fungsi kepadatan probabilitas seperti yang tertulis pada persamaan (2.1), maka* $\boldsymbol{Y}$ *berdistribusi normal multivariat dan dinotasikan* $\boldsymbol{Y}\sim N_p(\mu, \Sigma)$. *Jika p=1 maka indeks dihilangkan.*

2.1.5 Model Regresi Linear

Menurut Sembiring [9], model regresi adalah model yang memberikan gambaran mengenai hubungan antara variabel bebas dengan variabel tak bebas yang dipengaruhi oleh beberapa parameter regresi yang belum diketahui nilainya. Jika analisis regesi dilakukan untuk satu variabel bebas dengan satu variabel tak bebas, maka regresi ini dinamakan regresi linear sederhana dengan model $Y = \beta_0 + \beta_1 X + \varepsilon$. Jika $X_1, X_2, \ldots, X_p$ adalah variabel bebas dan Y adalah variabel tak bebas, maka regresi ini dinamakan regresi linear ganda dan model regresinya adalah $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ dengan $\varepsilon \sim N(0, \sigma^2)$.

2.1.6 Metode Kuadrat Terkecil

Sebagaimana ditulis Sembiring [9], pada model regresi linear sederhana dengan n data pengamatan (x_i, y_i), $i = 1, 2, \ldots, n$ akan ditentukan parameter re-

gresi β_0 dan β_1 sedemikian rupa sehingga

$$J = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.2}$$

minimum. Pada persamaan (2.2), x_i dan y_i bilangan yang berasal dari pengamatan sedangkan β_0 dan β_1 berubah bila garis regresinya berubah. Koefisien regresi β_0 dan β_1 dianggap berubah sehingga J diturunkan terhadap β_0 dan β_1 kemudian menyamakannya dengan 0, diperoleh

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad atau \quad \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \tag{2.3}$$

dan

$$\frac{\partial J}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \; atau \; \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2.4}$$

Jika nilai β_0 dan β_1 pada persamaan (2.3) dan (2.4) diganti dengan masing-masing estimatornya, $\hat{\beta}_0$ dan $\hat{\beta}_1$, maka persamaannya menjadi suatu sistem persamaan linear yang disebut dengan persamaan normal sebagai berikut

$$\begin{aligned} \sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i. \end{aligned} \tag{2.5}$$

Dari persamaan (2.5) yang pertama diperoleh

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

dengan $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ dan $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Persamaan (2.5) yang kedua menjadi

$$\sum_{i=1}^{n} y_i x_i - \{\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 (\frac{\sum_{i=1}^{n} x_i}{n})\} (\sum_{i=1}^{n} x_i) - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

atau

$$\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n} - \hat{\beta}_1 \{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\} = 0.$$

Jadi,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{(\sum_{i=1}^{n} x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}.$$

Rumus untuk $\hat{\beta}_1$ tersebut dapat dengan mudah disederhanakan menjadi

$$\begin{aligned}\hat{\beta}_1 &= \frac{\Sigma_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\Sigma_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= \frac{\Sigma_{i=1}^n x_i y_i - \bar{x}\Sigma_{i=1}^n y_i \bar{y}\Sigma_{i=1}^n x_i + n\bar{x}\bar{y}}{\Sigma_{i=1}^n x_i^2 - 2\bar{x}\Sigma_{i=1}^n x_i + n\bar{x}^2} \\ &= \frac{\Sigma_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^n (x_i - \bar{x})^2}\end{aligned} \tag{2.6}$$

Baris terakhir persamaan (2.6) dapat ditulis dengan

$$\hat{\beta}_1 = \Sigma\, w_i \,(\frac{y_i - \bar{y}}{x_i - \bar{x}}), \qquad w_i = \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$$

Notasi $(y_i - \bar{y})/(x_i - \bar{x}))$ merupakan *slope* garis antara titik pusat $(\bar{x}, \bar{y})$ dan titik data $(x_i,\, y_i)$. Jadi *slope* $\hat{\beta}_1$ pada estimasi garis regresi merupakan rata-rata dari beberapa *slope*. Dengan kata lain, $\hat{\beta}_1$ merupakan rata-rata terboboti. Bobotnya adalah w_i, dimana w_i harus nonnegatif dan jumlahnya sama dengan satu.

Perhatikan persamaan berikut

$$(y_i - \bar{y}) \equiv (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Bila ruas kiri dan kanan dikuadratkan dan kemudian dijumlahkan, maka diperoleh

$$\begin{aligned}\Sigma_{i=1}^n (y_i - \bar{y})^2 &\equiv \Sigma_{i=1}^n ((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i))^2 \\ &= \Sigma_{i=1}^n (\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^n (y_i - \hat{y}_i)^2 \\ &\quad + 2\,\Sigma_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i).\end{aligned} \tag{2.7}$$

Perhatikan perkalian silang persamaan (2.7)

$$\Sigma_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \Sigma_{i=1}^n \hat{y}_i (y_i - \hat{y}) - \bar{y}\Sigma_{i=1}^n (y_i - \hat{y}_i).$$

Berdasar persamaan normal (2.5) maka bagian kedua ruas kanan sama dengan nol,

$$\Sigma_{i=1}^n (y_i - \hat{y}_i) = \Sigma_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Bagian pertama ruas kanan ruas kanan juga sama dengan nol karena

$$\begin{aligned}\Sigma_{i=1}^n \hat{y}_i (y_i - \hat{y}) &= \Sigma_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i)(y_i - \hat{y}_i) \\ &= \hat{\beta}_0\, \Sigma_{i=1}^n (y_i - \hat{y}_i) + \hat{\beta}_1\, \Sigma_{i=1}^n (y_i - \hat{y}_i) x_i \\ &= 0 + \hat{\beta}_1\, \Sigma_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i \\ &= 0.\end{aligned}$$

Jadi persamaan (2.7) dapat ditulis dengan

$$\Sigma_{i=1}^{n}(y_i - \bar{y})^2 = \Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.8}$$

Persamaan (2.8) merupakan persamaan dasar dalam analisis regresi. Ruas kiri disebut jumlah kuadrat total (JKT) atau jumlah variasi total dan menyatakan jumlah penyimpangan y di sekitar rata-ratanya. Bagian pertama ruas kanan disebut jumlah kuadrat regresi (JKR). Bagian ini menyatakan pengaruh x terhadap y. Bagian kedua ruas kanan disebut jumlah kuadrat sisa (JKS), dimana bagian ini mengukur sisa dari variasi total yang tidak dapat diterangkan oleh x.

Model regresi linear yang mengandung p variabel bebas dapat dituliskan sebagai

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon. \tag{2.9}$$

Persamaan (2.9) disebut dengan model regresi linear ganda. Bila pengamatan terhadap variabel $Y, X_1, X_2, \ldots, X_p$ yang nilainya dinyatakan dengan $y_i, x_{i1}, x_{i2}, \ldots, x_{ip}$ dan sesatan random ε_i, maka persamaan (2.9) dapat dituliskan sebagai

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i\,, \qquad i = 1, 2, \ldots, n. \tag{2.10}$$

Persamaan (2.10) dapat dituliskan dalam lambang matriks sebagai berikut

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{2.11}$$

Jadi, persamaan (2.11) dapat ditulis sebagai persamaan

$$\boldsymbol{y} = \boldsymbol{X\beta} + \xi. \tag{2.12}$$

Pada persamaan (2.12), $\boldsymbol{y}$ menyatakan vektor variabel tak bebas n x 1, $\boldsymbol{X}$ menyatakan matriks variabel bebas n x $(p+1)$, $\boldsymbol{\beta}$ menyatakan vektor parameter $(p+1)$ x 1 dan ξ menyatakan vektor sesatan random n x 1. Bila vektor $\hat{y}$ adalah taksiran

dari $\boldsymbol{y}$, dan $\hat{ß}$ adalah taksiran dari $\boldsymbol{ß}$, maka taksiran kuadrat terkecil dari (2.12) dapat ditulis sebagai

$$\hat{y} = \boldsymbol{X}\hat{ß}.$$

Seperti pada model regresi linear, akan dicari vektor $\hat{ß} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ yang merupakan taksiran vektor parameter $\boldsymbol{ß} = (\beta_0, \beta_1, \ldots, \beta_p)'$, yang meminimumkan bentuk kuadrat

$$\begin{aligned} J &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{ß})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{ß}) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{ß} + ß'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{ß}. \end{aligned} \tag{2.13}$$

Pada persamaan (2.13), $\boldsymbol{y}'\boldsymbol{X}\boldsymbol{ß} = ß'\mathbf{X}'\boldsymbol{y}$ karena keduanya skalar. Berdasar persamaan (2.10) diperoleh

$$J = \Sigma\ \varepsilon_i^2 = \Sigma\ (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2.$$

Vektor $\hat{ß} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$ diperoleh dengan cara menurunkan J secara parsial terhadap $\beta_0, \beta_1, \ldots, \beta_p$ kemudian menyamakannya dengan nol, sehingga

$$\begin{aligned} \frac{\partial J}{\partial \beta_0} &= -2\Sigma\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) = 0 \\ \frac{\partial J}{\partial \beta_1} &= -2\Sigma\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) x_{i1} = 0 \\ \frac{\partial J}{\partial \beta_2} &= -2\Sigma\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) x_{i2} = 0 \\ &\vdots \\ \frac{\partial J}{\partial \beta_p} &= -2\Sigma\,(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}) x_{ip} = 0 \end{aligned} \tag{2.14}$$

Jika nilai $\beta_0, \beta_1, \ldots, \beta_p$ pada persamaan (2.14) diganti dengan masing-masing estimatornya, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, maka persamaannya menjadi suatu sistem persamaan

linear yang disebut dengan persamaan normal sebagai berikut

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} \\
\sum_{i=1}^{n} y_i x_{i1} &= \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}x_{i1} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip}x_{i1} \\
\sum_{i=1}^{n} y_i x_{i2} &= \hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip}x_{i2} \\
&\vdots \\
\sum_{i=1}^{n} y_i x_{ip} &= \hat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{ip} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}x_{ip} + \cdots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip}^2
\end{aligned} \tag{2.15}$$

Jika persamaan (2.15) ditulis dalam lambang matriks maka bentuknya menjadi

$$(\boldsymbol{X}'\boldsymbol{X})\hat{ß} = \boldsymbol{X}'\boldsymbol{y}. \tag{2.16}$$

Berdasar persamaan (2.16), maka diperoleh

$$\hat{ß} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

2.1.7 Uji Hipotesis

Definisi uji hipotesis di bawah ini diambil dari Walpole dan Myers [11].

Definisi 2.1.16. *Hipotesis statistik adalah suatu anggapan atau pernyataan, yang mungkin benar atau tidak, mengenai satu populasi atau lebih. Hipotesis ada dua macam, yaitu hipotesis nol dan hipotesis alternatif.*

Pengujian hipotesis terhadap suatu nilai parameter tergantung kasus yang diselidiki, akibatnya definisi terhadap kedua jenis hipotesis tersebut relatif terhadap kasus yang ada. Misal H_0 menyatakan hipotesis nol. Suatu nilai statistik uji yang diperoleh dari pengamatan dikatakan berarti jika H_0 ditolak pada taraf signifikansi α yang telah ditentukan. Himpunan nilai yang membuat penolakan H_0 disebut daerah kritis atau daerah penolakan. Secara umum, langkah-langkah uji hipotesis adalah

1. menentukan hipotesis nol H_0 dan hipotesis alternatif H_1,

2. memilih tingkat signifikansi α,

3. menentukan daerah kritis,

4. menghitung statistik uji,

5. mengambil kesimpulan.

Untuk membandingkan dua perlakuan maka diperlukan suatu uji hipotesis yang disebut dengan uji dua sampel berpasangan. Misal X dan Y berkaitan dengan perlakuan 1 dan 2. Dua sampel yang berpasangan diartikan sebagai sebuah sampel dengan subjek yang sama namun mendapatkan dua perlakuan yang berbeda. Pasangan (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) independen dan diasumsikan perbedaan perlakuan yaitu $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$ berdistribusi normal dengan mean $= \delta$ dan variansi $= \sigma_D^2$. Walaupun pasangan (X_i, Y_i) independen, namun X_i dan Y_i dependen pada setiap pasangan ke-i (Johnson dan Bhattacharyya [5]). Jika $\delta = 0$ berarti kedua perlakuan sama, $\delta > 0$ berarti mean perlakuan 1 lebih besar dari mean perlakuan 2 dan $\delta < 0$ berarti mean perlakuan 1 lebih kecil dari mean perlakuan 2. Mean sampel dan variansi sampel untuk D adalah $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ dan $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$.

Uji hipotesis untuk $H_0 : \mu_1 - \mu_2 = \delta_0$ berdasar pada statistik uji

$$T = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}}$$

dengan derajat bebas $n - 1$. Penentuan daerah penolakan H_0 adalah

1. untuk $H_1 : \mu_1 - \mu_2 \neq 0$, H_0 ditolak jika $|t| > t_{\alpha/2}$,

2. untuk $H_1 : \mu_1 - \mu_2 > 0$, H_0 ditolak jika $t > t_\alpha$,

3. untuk $H_1 : \mu_1 - \mu_2 < 0$, H_0 ditolak jika $t < -t_\alpha$.

Untuk n besar, $n > 30$, digunakan distribusi normal, sehingga $\frac{\bar{D}-\delta_0}{S_D/\sqrt{n}} \sim N(0, 1)$.

2.1.8 Probabilitas Bersyarat dan Fungsi Likelihood

Pengertian tentang probabilitas bersyarat peristiwa A jika diketahui B telah terjadi yang ditulis sebagai $P(A|B)$, dua kejadian independen dan fungsi likelihood diberikan pada definisi dan teorema berikut yang diambil dari Bain dan Engelhardt [2].

Definisi 2.1.17. *Misal dipunyai ruang sampel S dan A,B adalah peristiwa di dalam S. Probabilitas A dengan syarat B didefinisikan sebagai*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{dengan } P(B) \neq 0.$$

Definisi 2.1.18. *Dua kejadian A dan B disebut kejadian independen jika*

$$P(A \cap B) = P(A)P(B)$$

selain itu A dan B disebut kejadian dependen.

Definisi 2.1.19. *Fungsi kepadatan probabilitas bersama dari n variabel random $T_1, T_2, \ldots, T_n$ yang diberi nilai $t_1, t_2, \ldots, t_n$ dinotasikan dengan $f(t_1, t_2, \ldots, t_n; \theta)$ merupakan fungsi likelihood. Untuk nilai $t_1, t_2, \ldots, t_n$ tertentu, fungsi likelihoodnya merupakan fungsi dari θ dan dinotasikan dengan $L(\theta)$. Jika $T_1, T_2, \ldots, T_n$ adalah sampel random yang independen maka*

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta) = f(t_1; \theta) f(t_2; \theta) \ldots f(t_n; \theta)$$

dengan θ adalah parameter yang tidak diketahui.

2.1.9 Distribusi *Prior* dan Distribusi *Posterior*

Definisi distribusi *prior* dan distribusi *posterior* di bawah ini diambil dari Larson [6].

Definisi 2.1.20. *Distribusi prior dari suatu parameter θ merupakan fungsi kepadatan probabilitas yang menggambarkan tingkat keyakinan nilai θ.*

Sebagaimana ditulis oleh Larson [6], distribusi *prior* tersebut diperoleh sebelum melakukan analisis data.

Definisi 2.1.21. *Fungsi kepadatan posterior untuk θ merupakan fungsi kepadatan probabilitas bersyarat θ diberikan nilai sampel y, sehingga*

$$f(\theta|y) = \frac{f(y,\theta)}{f(y)}\,.$$

Secara umum, distribusi *prior* menggambarkan tingkat keyakinan terhadap kemungkinan nilai parameter θ sedangkan distribusi *posterior* menggambarkan tingkat keyakinan terhadap kemungkinan nilai parameter θ setelah diberikan nilai sampel.

2.1.10 Teorema Bayes

Percobaan awal yang telah dilakukan akan berpengaruh terhadap hasil percobaan sekarang. Untuk menghitung probabilitas kejadian sekarang dengan syarat kejadian awal dijelaskan dalam teorema Bayes. Teorema 2.1.1 di bawah ini diambil dari Bain dan Engelhardt [2], sedangkan Teorema 2.1.2 dan Teorema 2.1.3 diambil dari Walpole dan Myers [11].

Teorema 2.1.1. *Jika A dan B adalah dua kejadian sembarang, maka*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Akibatnya, jika A dan B adalah kejadian yang saling asing, maka $A \cap B = \emptyset$, sehingga $P(A \cup B) = P(A) + P(B)$. Selanjutnya, jika $A_1, A_2, \ldots, A_n$ saling asing, maka $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$.

Teorema 2.1.2. *Misal kejadian $B_1, B_2, \ldots, B_k$ merupakan kejadian yang saling asing dari ruang sampel S dengan $P(B_i) \neq 0$ untuk $i = 1, 2, \ldots, k$, maka untuk setiap kejadian A anggota S*

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

Teorema 2.1.3. *Misal kejadian $B_1, B_2, \ldots, B_k$ merupakan kejadian yang saling asing dari ruang sampel S dengan $P(B_i) \neq 0$ untuk $i = 1, 2, \ldots, k$. Misalkan A suatu kejadian sembarang dalam S dengan $P(A) \neq 0$, maka*

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

untuk $r = 1, 2, \ldots, k$.

Teorema 2.1.3 disebut sebagai teorema Bayes. Teorema Bayes memberikan aturan untuk menghitung probabilitas bersyarat peristiwa B_r diberikan A, jika masing-masing probabilitas tak bersyarat B_r dan probabilitas bersyarat A diberikan B_r diketahui. Untuk selanjutnya $P(B_r)$ disebut probabilitas *prior*, $P(A|B_r)$ disebut sebagai fungsi likelihood dan $P(B_r|A)$ disebut fungsi probabilitas *posterior*.

2.1.11 Estimator Bayes

Estimasi parameter berdasar pada prinsip Bayesian akan selalu berhubungan dengan distribusi *prior* dan distribusi *posterior*. Berikut diberikan definisi tentang estimator Bayes yang diambil dari Larson [6].

Definisi 2.1.22. *Harga harapan distribusi posterior* θ *disebut estimator Bayes untuk* θ.

2.2 Kerangka Pemikiran

Kerangka pemikiran dalam penulisan skripsi ini dijelaskan sebagai berikut. Analisis regresi merupakan teknik statistik untuk menyelidiki hubungan antar variabel. Metode yang biasa digunakan untuk mengestimasi parameter regresi adalah metode kuadrat terkecil. Pada prakteknya, saat melakukan inferensi tentang parameter populasi kadang-kadang diperoleh informasi tambahan mengenai parameter populasi, dimana informasi tersebut berasal dari data awal. Jika informasi tersebut ingin dimasukkan dalam analisis data, maka estimasi dengan metode kuadrat terkecil tidak dapat digunakan. Oleh karena itu, diperlukan metode Bayesian untuk menyelesaikan permasalahan tersebut. Secara praktis, analisis dengan metode Bayesian memerlukan distribusi *prior* dan distribusi *posterior*. Ada dua macam distribusi *prior*, yaitu distribusi *prior* noninformatif dan distribusi *prior* sekawan. Harga harapan distribusi *posterior* digunakan untuk mengestimasi parameter dalam model regresi. Setelah diperoleh estimasi parame-

ter, dilakukan uji signifikansi terhadap parameter tersebut untuk mengetahui bagaimana pengaruh variabel bebas terhadap variabel tak bebas. Selanjutnya, dilakukan analisis eror untuk membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

BAB III

METODE PENELITIAN

Metode yang ditempuh dalam penulisan skripsi ini adalah studi literatur yaitu dengan mengumpulkan referensi berupa buku-buku yang dapat mendukung pembahasan mengenai estimasi parameter regresi linear dengan metode Bayesian, sedangkan untuk perhitungan pada contoh kasus menggunakan bantuan *sofware SPSS 11 for Windows*, *Minitab 11 for Windows* dan *Microsoft Excel*.

Adapun langkah-langkah yang ditempuh dalam penulisan ini adalah

1. menjelaskan metode Bayesian,
2. mengumpulkan apa yang diketahui tentang kemungkinan nilai parameter, hal inilah yang menjadi bentuk distribusi *prior* untuk parameter tersebut, selanjutnya distribusi *prior* digunakan untuk memperoleh distribusi *posterior*,
3. mengestimasi parameter regresi linear melalui harga harapan distribusi *posterior*,
4. menentukan probabilitas *posterior* di bawah hipotesis nol untuk menguji signifikansi parameter regresi berdasar distribusi *prior* noninformatif,
5. menghitung eror model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan,
6. nilai eror yang telah diperoleh pada langkah lima digunakan untuk membandingkan estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan pada suatu kasus.

BAB IV

PEMBAHASAN

Sebagaimana ditulis oleh Birkes dan Dodge [3], data yang terdiri dari n observasi dengan satu variabel tak bebas Y dan p variabel bebas X_1, X_2, ..., X_p disebut data regresi karena pada umumnya digunakan suatu model regresi untuk menganalisisnya. Model regresi linear adalah $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$, sedangkan untuk sebuah data observasi model regresinya adalah $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$ dengan $i = 1, 2, \ldots, n$ dan sesatan random ε_i merupakan penyimpangan model dari keadaan sesungguhnya. Diasumsikan ε_i berbentuk acak dan harga harapannya nol. Menganggap ε_i acak berarti bahwa sampel diambil secara acak. Jika model regresi terdiri dari satu variabel bebas dan satu variabel tak bebas maka dinamakan model regresi linear sederhana dan modelnya adalah $Y = \beta_0 + \beta_1 x_1 + \varepsilon$.

Jika diketahui informasi *prior* tentang data yang akan dianalisis, maka informasi tersebut dapat dimasukkan dalam analisis dengan menggunakan metode Bayesian. Pada bab ini dibahas mengenai metode Bayesian, estimasi parameter regresi linear sederhana dan regresi linear ganda, uji signifikansi parameter regresi, serta perbandingan estimasi model regresi berdasar distribusi *prior* sekawan dan distribusi *prior* noninformatif pada suatu kasus.

4.1 Metode Bayesian

Estimasi parameter dan pengujian hipotesis biasanya dilakukan berdasar pada informasi sampel yang diambil dari populasi. Namun pada prakteknya, ketika melakukan inferensi tentang parameter populasi kadang-kadang diperoleh informasi tambahan mengenai parameter populasi, dimana informasi tersebut berasal dari data sebelumnya. Metode Bayesian merupakan suatu metode un-

tuk menghasilkan estimasi parameter dan uji hipotesis dengan menggabungkan informasi dari sampel (data sekarang) dan informasi lain yang telah tersedia sebelumnya (Birkes dan Dodge [3], Soejoeti dan Soebanar [10]).

Sebelum mengaplikasikan metode Bayesian terhadap masalah regresi, terlebih dahulu akan diberikan garis besar mengenai ciri-ciri umumnya. Misalkan $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ menyatakan vektor data dan $\theta = (\beta_0, \beta_1, \sigma)$ menyatakan vektor parameter yang tidak diketahui dari model regresi linear sederhana. Data $\boldsymbol{y}$ digunakan untuk mengestimasi maupun menguji hipotesis beberapa parameter dalam θ.

Vektor data $\boldsymbol{y}$ merupakan variabel random yang mempunyai distribusi probabilitas tertentu untuk setiap nilai tetap vektor parameter θ. Model regresi linear sederhana menjelaskan bahwa untuk nilai tetap β_0, β_1, σ maka y_i berdistribusi normal dengan mean $\beta_0 + \beta_1 x_i$, variansi σ^2 dan y_i saling independen. Distribusi $\boldsymbol{y}$ untuk nilai tetap θ disebut distribusi bersyarat $\boldsymbol{y}$ diberikan nilai θ. Pada metode Bayesian, harus ditentukan distribusi untuk θ. Sebelum menganalisis dan memberikan interpretasi tentang data, terlebih dahulu harus menaksir apa yang diketahui tentang kemungkinan nilai parameter, hal inilah yang menjadi bentuk distribusi probabilitas θ, distribusi ini disebut distribusi *prior* θ. Langkah selanjutnya adalah mengkombinasikan distribusi *prior* θ dan distribusi bersyarat $\boldsymbol{y}$ diberikan nilai θ untuk mendapatkan distribusi bersyarat θ diberikan $\boldsymbol{y}$. Untuk vektor data $\boldsymbol{y}$ yang diobservasi, distribusi bersyarat θ diberikan $\boldsymbol{y}$ disebut dengan distribusi *posterior*.

Rumus Bayes merupakan alat untuk mengkombinasikan distribusi θ dan distribusi $\boldsymbol{y}$ dengan syarat θ untuk mendapatkan distribusi θ dengan syarat $\boldsymbol{y}$. Distribusi ini sering disajikan dalam bentuk fungsi kepadatan probabilitas $f(\theta)$, $f(\boldsymbol{y}|\theta)$ dan $f(\theta|\boldsymbol{y})$. Sebagaimana ditulis oleh Birkes dan Dodge [3], rumus Bayes adalah

$$f(\theta|\boldsymbol{y}) = C f(\theta) f(\boldsymbol{y}|\theta) \tag{4.1}$$

dengan C merupakan konstanta yang tidak bergantung pada θ, yaitu

$$C = \frac{1}{\int_{-\infty}^{\infty} f(\theta) f(\boldsymbol{y}|\theta) d\theta} \, .$$

Estimasi Bayes maupun uji hipotesis tentang parameter dalam θ diperoleh dari distribusi *posterior* θ. Sebagai contoh, θ dapat diestimasi dengan harga harapan distribusi *posterior*.

4.2 Estimasi Parameter Regresi Linear Sederhana

Pertama-tama akan ditentukan distribusi probabilitas vektor data. Diasumsikan model regresi linear normal, sehingga untuk nilai tetap β_0, β_1, σ, vektor data $\boldsymbol{y}= (y_1, y_2, \ldots, y_n)$ berdistribusi normal multivariat.

Metode Bayesian memerlukan distribusi *prior* bersama untuk parameter β_0, β_1, σ. Secara teoritis, sembarang distribusi dapat dipilih sebagai distribusi *prior*. Tentu saja, distribusi *posterior* bergantung pada distribusi *prior* dan fungsi likelihood atau distribusi probabilitas vektor data. Namun pada prakteknya, perhitungan untuk distribusi *posterior* akan lebih mudah dilakukan jika dipilih distribusi *prior* yang dapat dikombinasikan dengan baik dengan distribusi vektor data yang dipunyai. Dengan perkataan lain, dipilih distribusi *prior* dalam keluarga distribusi tertentu yang tergantung pada bentuk fungsi likelihoodnya. Caranya adalah dengan memilih $f(\theta)$ yang mempunyai struktur yang sama dengan $f(y|\theta)$, dalam hal ini $f(\theta)$ sekawan dengan likelihood sampel. Karena $f(\theta)$ dan $f(y|\theta)$ dalam keluarga yang sama maka $f(\theta|y)$ akan mempunyai fungsi kepadatan probabilitas yang sama dengan $f(\theta)$. Berikut disajikan dua tipe distribusi *prior*, yaitu distribusi *prior* noninformatif dan distribusi *prior* sekawan.

4.2.1 Distribusi *Prior* Noninformatif Regresi Linear Sederhana

Menurut Soejoeti dan Soebanar [10], distribusi *prior* harus mencerminkan informasi *prior* dari seorang peneliti, karena peneliti yang berbeda biasanya akan mempunyai informasi yang berbeda pula. Diasumsikan informasi *prior* atau ide mengenai kemungkinan nilai β_0, β_1, σ tidak diketahui, melainkan σ harus positif, sehingga diperlukan distribusi *prior* noninformatif yang menunjukkan ketidak-

tahuan tentang kemungkinan nilai parameter. Fungsi kepadatan probabilitas distribusi *prior* noninformatif adalah

$$f(\beta_0, \beta_1, \sigma) = 1/\sigma. \tag{4.2}$$

Persamaan (4.2) menjelaskan bahwa sebelum mengolah data dan memberikan interpretasinya, semua nilai β_0 mungkin sama, semua nilai β_1 mungkin sama dan semua nilai σ mungkin sama. Persamaan (4.2) disebut dengan distribusi *prior* standar.

Menurut Birkes dan Dodge [3], estimasi Bayes yang dihasilkan sama dengan estimasi kuadrat terkecil, sehingga dengan menggunakan Definisi 2.1.22 diperoleh estimator Bayes untuk β_0 dan β_1 sebagai berikut

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4.3}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{(\sum_{i=1}^{n} x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{4.4}$$

dengan $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ dan $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.
Walaupun estimasi dengan metode Bayesian dan metode kuadrat terkecil memberikan hasil yang sama, tapi alasan keduanya sangat berbeda.

4.2.2 Distribusi *Prior* Sekawan Regresi Linear Sederhana

Misalkan informasi *prior* tentang parameter diketahui, maka informasi tersebut dinyatakan dalam bentuk distribusi *prior* sekawan (Birkes dan Dodge [3]). Kombinasi distribusi *prior* sekawan dengan distribusi vektor data menghasilkan distribusi *posterior* yang mempunyai bentuk distribusi yang sama dengan distribusi *prior*. Menurut Birkes dan Dodge [3], vektor data berdistribusi normal multivariat berarti bila distribusi (β_0, β_1) dengan syarat σ merupakan distribusi normal bivariat dan distribusi $1/\sigma^2$ merupakan distribusi gamma, maka distribusi (β_0, β_1) dengan syarat σ dan distribusi $1/\sigma^2$ merupakan distribusi *prior* sekawan untuk parameter $(\beta_0, \beta_1, \sigma)$.

Ada kesulitan yang mungkin dihadapi dalam penggunaan persamaan (4.1). Jika $f(\theta)$ dan $f(\boldsymbol{y}|\theta)$ bukan merupakan fungsi matematika yang sederhana, maka pengintegralan penyebutnya akan sulit dilakukan. Cara untuk menghindari kesulitan ini adalah dengan membatasi diri pada distribusi *prior* dalam keluarga distribusi tertentu yang tergantung pada bentuk fungsi likelihoodnya. Distribusi *prior* sekawan merupakan keluarga distribusi yang memudahkan hitungan bila digunakan sebagai distribusi *prior*. Distribusi tersebut dapat berupa ringkasan angka dari data sebelumnya (Birkes dan Dodge [3]). Informasi *prior* bisa diperoleh dengan menentukan harga harapan, deviasi standar serta korelasi antara β_0 dan β_1.

Misalkan $\mu = \beta_0 + \beta_1 x_m$ dengan x_m merupakan nilai rata-rata x pada data sekarang maka persamaan tersebut dapat ditulis kembali menjadi $\mu = \beta_0 + \beta_1 \bar{x}$. Berdasar informasi *prior*, ditentukan distribusi *prior* sekawan yang meliputi harga harapan μ dan β yaitu e_μ dan e_β serta deviasi standar μ dan β yaitu c_μ dan c_β. Dengan menggunakan Definisi 2.1.22 diperoleh estimator Bayes untuk β_0 dan β_1 sebagai berikut

$$\begin{aligned} \hat{\beta}_0 &= \hat{\mu} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \{\frac{c_{\beta_1}^{-2}}{c_{\beta_1}^{-2} + \sum(x_i - \bar{x}^2)}\} e_{\beta_1} + \{\frac{\sum(x_i - \bar{x}^2)}{c_{\beta_1}^{-2} + \sum(x_i - \bar{x}^2)}\} (\hat{\beta}_1)_{LS} \end{aligned} \tag{4.5}$$

dengan

$$\hat{\mu} = \{\frac{c_\mu^{-2}}{c_\mu^{-2} + n}\} e_\mu + \{\frac{n}{c_\mu^{-2} + n}\} \bar{y}$$

dan $(\hat{\beta}_1)_{LS}$ merupakan estimasi kuadrat terkecil untuk data sekarang (Birkes dan Dodge [3]).

Persamaan (4.5) menunjukkan informasi *prior* yang digabung dengan informasi dari data sekarang. Estimator Bayes untuk β_1 merupakan rata-rata tertimbang dari e_{β_1} (harga harapan *prior* β_1 berdasar informasi awal) dan $(\hat{\beta}_1)_{LS}$ (estimasi kuadrat terkecil β_1 berdasar data sekarang). Estimator Bayes untuk μ juga merupakan rata-rata tertimbang dari harga harapan *prior* berdasar informasi awal dan estimasi kuadrat terkecil berdasar data sekarang, dalam hal ini $\bar{y} = \hat{\mu}_{LS}$.

4.3 Uji Signifikansi Parameter Regresi Linear Sederhana

Uji hipotesis pada metode Bayesian dilakukan dengan menghitung probabilitas distribusi *posterior* bahwa hipotesis yang dibuat adalah benar. Probabilitas *posterior* H_0 adalah Prob$(H_0|D)$ dengan D menyatakan kejadian bahwa data dalam percobaan yang diulang sama dengan data yang sebenarnya.

Misalkan informasi *prior* tentang parameter tidak diketahui. Selanjutnya akan dirumuskan distribusi *prior* yang sesuai untuk keadaan tersebut. Distribusi *prior* untuk hipotesis nol, $H_0 : \beta_1 = 0$, dan hipotesis alternatif, $H_1 : \beta_1 \neq 0$, adalah

$$Prob(H_0) = \frac{1}{2}\,, \qquad Prob(H_1) = \frac{1}{2} \tag{4.6}$$

Probabilitas *prior* hipotesis nol dan hipotesis alternatif pada persamaan (4.6) bernilai sama disebabkan karena tidak diketahuinya informasi *prior* tentang parameter.

Uji Bayesian untuk $\beta_1 = 0$ dapat ditunjukkan dengan menghitung probabilitas *posterior* hipotesis nol, yaitu

$$Prob(H_0|y) = \frac{1}{1 + \frac{1}{\sqrt{g}}} \tag{4.7}$$

dengan

$$g = (n+1)[1 - (\frac{n}{n+1})r^2]^{n-1}$$

dan r merupakan koefisien korelasi sampel antara x dan y

$$r = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}\,.$$

Daerah penolakan untuk hipotesis $\beta_1 = 0$ pada uji Bayesian adalah Prob$(H_0|y)$ kurang dari Prob$(H_0) = \frac{1}{2}$. Nilai $|r|$ yang besar mengindikasikan adanya hubungan yang kuat antara x dan y. Hal ini berarti $\beta_1 \neq 0$ sehingga probabilitas *posterior* untuk H_0 harus kecil agar didapatkan hubungan tersebut.

4.4 Estimasi Parameter Regresi Linear Ganda

Model regresi yang terdiri dari satu variabel tak bebas dan lebih dari satu variabel bebas disebut model regresi linear ganda. Misalkan $\boldsymbol{y}$ menyatakan vek-

tor variabel tak bebas, $\boldsymbol{X}$ merupakan matriks variabel bebas dan $\boldsymbol{ß}$ merupakan vektor parameter regresi. Untuk nilai tetap parameter $\boldsymbol{ß}$ dan σ, $\boldsymbol{y}$ diasumsikan berdistribusi normal multivariat. Seperti yang telah dijelaskan pada subbab 4.2, berikut disajikan dua tipe distribusi *prior* untuk parameter, yaitu distribusi *prior* noninformatif dan distribusi *prior* sekawan.

4.4.1 Distribusi *Prior* Noninformatif Regresi Linear Ganda

Misalkan informasi *prior* tentang parameter tidak diketahui. Seperti yang dijelaskan di 4.2.1, distribusi *prior* yang bisa digunakan untuk menunjukkan ketidaktahuan tentang kemungkinan nilai parameter diberikan oleh fungsi kepadatan

$$f(\boldsymbol{ß}, \sigma) = 1/\sigma.$$

Menurut Birkes dan Dodge [3], estimasi Bayes yang dihasilkan sama dengan estimasi kuadrat terkecil, sehingga dengan menggunakan Definisi 2.1.22 diperoleh estimator Bayes untuk vektor parameter $\boldsymbol{ß}$ sebagai berikut

$$\hat{ß} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{4.8}$$

Rumus Bayes digunakan untuk mengetahui mengapa estimasi Bayes dengan distribusi *prior* noninformatif sama dengan estimasi kuadrat terkecil. Rumus tersebut biasa digunakan untuk memperoleh distribusi *posterior* untuk semua parameter β dan σ. Fungsi kepadatan distribusi *posterior* dinyatakan dengan $f(\beta|\boldsymbol{y}, \sigma)$. Pada persamaan (4.1), posisi θ digantikan oleh $\boldsymbol{ß}$ dan semua distribusi bersyarat pada σ, sehingga

$$f(\boldsymbol{ß}|\boldsymbol{y}, \sigma) = Cf(\boldsymbol{ß}|\sigma)f(\boldsymbol{y}|\boldsymbol{ß}, \sigma) \tag{4.9}$$

dengan C merupakan konstanta yang tidak bergantung pada $\boldsymbol{ß}$.

Menurut Soejoeti dan Soebanar [10], misal akan diperkirakan distribusi *prior* dalam keadaan hanya dipunyai informasi yang sangat sedikit atau tidak dipunyai informasi sama sekali. Secara lebih spesifik, dapat dikatakan informasi *prior*

tersebut dapat diwakili oleh informasi sampel. Keadaan yang digambarkan itu tidak harus merupakan keadaan informasi dalam arti sebenarnya, melainkan tanpa informasi dalam arti relatif. Jika distribusi *prior* dapat diwakili oleh informasi sampel, maka distribusi *prior* hampir seluruhnya bergantung pada fungsi likelihood, sehingga $f(\boldsymbol{ß}|\boldsymbol{y},\sigma) = Cf(\boldsymbol{y}|\boldsymbol{ß},\sigma)$.

Diasumsikan $\boldsymbol{y}$ berdistribusi normal multivariat dengan mean vektor $\boldsymbol{Xß}$ dan matriks variansi kovariansi $\sigma^2\boldsymbol{I}$, dimana $\boldsymbol{I}$ menyatakan matriks identitas. Fungsi kepadatan probabilitas distribusi normal multivariat adalah

$$f(\boldsymbol{y}|\boldsymbol{ß},\sigma) = C\exp[-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{Xß}\|^2]. \tag{4.10}$$

Pada metode kuadrat terkecil, vektor residual $\boldsymbol{y}$-$\boldsymbol{Xß}$ tegak lurus terhadap semua kolom matriks regresi $\boldsymbol{X}$, yang secara tidak langsung menyatakan bahwa $\| \boldsymbol{y}-\boldsymbol{Xß} \|^2 = \| \boldsymbol{y}-\boldsymbol{X}\hat{ß}_{LS} \|^2 + \| \boldsymbol{X}\hat{ß}_{LS}-\boldsymbol{Xß} \|^2$, sehingga

$$\begin{aligned} f(\beta|\boldsymbol{y},\sigma) &= C\exp[-\frac{1}{2\sigma^2} \| \boldsymbol{X}\hat{ß}_{LS}-\boldsymbol{Xß} \|^2] \\ &= C\exp[-\frac{1}{2\sigma^2}(\boldsymbol{ß}-\hat{ß}_{LS})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{ß}-\hat{ß}_{LS})]. \end{aligned} \tag{4.11}$$

Pada persamaan (4.11) terlihat bahwa fungsi kepadatan probabilitas distribusi *posterior* $\boldsymbol{ß}$ dengan syarat σ adalah fungsi kepadatan probabilitas distribusi normal multivariat dengan mean vektor $\hat{ß}_{LS}$ dan matriks variansi kovariasi $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

4.4.2 Distribusi *Prior* Sekawan Regresi Linear Ganda

Menurut Birkes dan Dodge [3], ukuran sampel kecil mungkin menyebabkan estimasi regresi menjadi kurang tepat, sehingga diperlukan informasi tambahan yang berasal dari data awal dimana informasi tersebut berpengaruh terhadap terhadap peningkatan ketelitian estimasi regresi. Metode Bayesian digunakan untuk memasukkan informasi tersebut ke dalam analisis data.

Informasi dari data awal dijelaskan dalam bentuk distribusi *prior* untuk parameter. Untuk tujuan mengestimasi $\boldsymbol{ß}$, tidak perlu ditentukan semua distribusi *prior* $\boldsymbol{ß}$ dan σ tetapi cukup ditentukan distribusi *prior* $\boldsymbol{ß}$ dengan syarat σ. Persamaan (4.9) digunakan untuk mendapatkan fungsi kepadatan probabili-

tas *posterior* $\boldsymbol{ß}$ dengan syarat σ yaitu dengan menggabungkan fungsi kepadatan probabilitas *prior* $f(\boldsymbol{ß}|\sigma)$ dan fungsi kepadatan probabilitas vektor data pada persamaan (4.10). Berdasar hal tersebut, maka dipilih fungsi kepadatan probabilitas *prior* yang mempunyai bentuk yang sesuai apabila dikombinasikan dengan persamaan (4.10). Bentuk distribusi *prior* $\boldsymbol{ß}$ dengan syarat σ adalah distribusi normal multivariat, yaitu

$$f(\boldsymbol{ß}|\sigma) = C \exp[-\frac{1}{2\sigma^2}(\boldsymbol{ß} - \boldsymbol{b})' \boldsymbol{V}^{-1}(\boldsymbol{ß} - \boldsymbol{b})] \tag{4.12}$$

dengan $\boldsymbol{b}$ adalah vektor mean dan $\sigma^2 \boldsymbol{V}$ adalah matriks variansi kovariansi. Vektor mean $\boldsymbol{b}$ dan matriks $\boldsymbol{V}$ dipilih untuk menggambarkan pengetahuan *prior*.

Berdasarkan informasi *prior*, akan ditentukan vektor mean dan matriks variansi kovariansi untuk distribusi *prior* $\boldsymbol{ß}$. Estimasi $\boldsymbol{ß}$ yang didapat dari data awal dipilih sebagai vektor mean $\boldsymbol{b}$, sedangkan estimasi matriks variansi kovariansi $\boldsymbol{ß}$ dari data awal dipilih sebagai $\boldsymbol{V}$ untuk matriks variansi kovariansi $\sigma^2 \boldsymbol{V}$.

Dengan menggunakan Definisi 2.1.22 diperoleh estimator Bayes untuk vektor parameter $\boldsymbol{ß}$ sebagai berikut

$$\hat{ß}_{Bayes} = \boldsymbol{V}_* \boldsymbol{V}^{-1} \boldsymbol{b} + \boldsymbol{V}_* \boldsymbol{X}' \boldsymbol{X} \hat{ß}_{LS} \tag{4.13}$$

dengan

$$\boldsymbol{V}_* = (\boldsymbol{V}^{-1} + \boldsymbol{X}' \boldsymbol{X})^{-1}.$$

Persamaan (4.13) merupakan rata-rata tertimbang dari harga harapan *prior* $\boldsymbol{ß}$ dan estimasi kuadrat terkecil $\boldsymbol{ß}$, bobotnya adalah $\boldsymbol{V}_* \boldsymbol{V}^{-1}$ dan $\boldsymbol{V}_* \boldsymbol{X}' \boldsymbol{X}$ (Birkes dan Dodge [3]). Bobot tersebut bila dijumlahkan maka hasilnya sama dengan matriks identitas.

4.5 Uji Signifikansi Parameter Regresi Linear Ganda

Model yang baik hendaknya selalu memperhatikan kesederhanaan dan keefektifan model. Menurut Birkes dan Dodge [3], prinsip kesederhanaan pada analisis regresi linear adalah mengambil satu atau lebih variabel bebas $X_{q+1}, \ldots, X_p$ dari model regresi keseluruhan $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ untuk mendapatkan

model regresi linear tereduksi $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \varepsilon$, dengan $q < p$. Akan diuji apakah terdapat perbedaan yang signifikan antara model keseluruhan dan model tereduksi. Jika tidak terdapat perbedaan yang signifikan, maka kedua model tersebut ekuivalen dan dipilih model tereduksi. Membandingkan dua model tersebut berarti menguji $\beta_{q+1} = \cdots = \beta_p = 0$.

Seperti pada model regresi linear sederhana, uji hipotesis metode Bayesian untuk model regresi linear ganda dengan p variabel bebas juga dilakukan dengan menghitung pobabilitas *posterior* bahwa hipotesis yang dibuat adalah benar.

Misal informasi *prior* tentang parameter tidak diketahui. Selanjutnya, dimisalkan $\boldsymbol{W}$ adalah matriks variabel bebas $x_{i1}, \ldots, x_{iq}$ dengan ukuran n x $(q+1)$ serta $\boldsymbol{Z}$ adalah matriks variabel bebas $x_{i,q+1}, \ldots, x_{ip}$ dengan ukuran n x $(p-q)$, jadi $\boldsymbol{X = (W, Z)}$. Misal $\gamma = (\beta_0, \ldots, \beta_q)$ dan $\delta = (\beta_{q+1}, \ldots, \beta_p)$, maka model $\boldsymbol{y = X\beta +}\ \xi$ dapat ditulis menjadi

$$\boldsymbol{y} = \boldsymbol{W}\gamma + \boldsymbol{Z}\delta + \xi.$$

Uji hipotesisnya adalah δ=0. Distribusi *prior* untuk hipotesis nol, $H_0 : \delta = 0$, dan hipotesis alternatif, $H_1 : \delta \neq 0$ adalah

$$Prob(H_0) = \frac{1}{2}, \qquad Prob(H_1) = \frac{1}{2} \tag{4.14}$$

Probabilitas *prior* hipotesis nol dan hipotesis alternatif pada persamaan (4.14) bernilai sama disebabkan karena tidak diketahuinya informasi *prior* tentang parameter.

Uji Bayesian untuk $\beta_{q+1} = \cdots = \beta_p = 0$ ditunjukkan dengan menghitung probabilitas *posterior* hipotesis nol, yaitu

$$Prob(H_0|\boldsymbol{y}) = \frac{1}{1 + \frac{1}{\sqrt{g}}} \tag{4.15}$$

dengan

$$g = (n+1)^{p-q}[1 - (\frac{n}{n+1})R^2]^{n-q-1}$$
$$R^2 = \frac{JKS_{tereduksi} - JKS_{keseluruhan}}{JKS_{tereduksi}}.$$

$JKS_{keseluruhan}$ adalah jumlah kuadrat sisa dengan metode kuadrat terkecil pada

model keseluruhan $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$, dan $JKS_{tereduksi}$ adalah jumlah kuadrat sisa dengan metode kuadrat terkecil pada model tereduksi $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \varepsilon$. Probabilitas *posterior* juga bisa diperoleh dengan terlebih dahulu menenetukan statistik uji dalam metode kuadrat terkecil yaitu F_{LS} karena

$$R^2 = \frac{JKR}{JKT} = \frac{JKR}{JKR + JKS} = \frac{JKR/JKS}{1 + JKR/JKS}$$

sedangkan

$$F_{LS} = \frac{JKR/p}{JKS/n - p - 1}$$

bila p menyatakan banyak variabel bebas dalam model, maka

$$\begin{aligned} R^2 &= \frac{p\,F_{LS}/(n-p-1)}{1 + p\,F_{LS}/(n-p-1)} \\ &= \frac{p\,F_{LS}}{(n-p-1) + (p\,F_{LS})} \\ &= \frac{F_{LS}}{F_{LS} + (n-p)/p} \\ &= \frac{1}{1 + (\frac{n-p-1}{p})\frac{1}{F_{LS}}} \\ &= \frac{1}{1 + (\frac{n-p-1}{p-q})\frac{1}{F_{LS}}} \end{aligned} \tag{4.16}$$

dalam hal in diambil $q = 0$, dengan

$$F_{LS} = \frac{JKS_{tereduksi} - JKS_{keseluruhan}}{(p-q)\hat{\sigma}^2}$$

dan

$$\hat{\sigma}^2 = \frac{JKS_{keseluruhan}}{n-p-1}\,.$$

Derah penolakan untuk hipotesis $\delta = 0$ pada uji Bayesian adalah Prob$(H_0|\boldsymbol{y})$ kurang dari Prob$(H_0) = \frac{1}{2}$.

4.6 Contoh Kasus

4.6.1 Contoh Kasus Regresi Linear Sederhana

Pola cuaca di utara United States menyebabkan curah hujan di Seattle berhubungan dengan curah hujan di Portland (Birkes dan Dodge [3]). Untuk

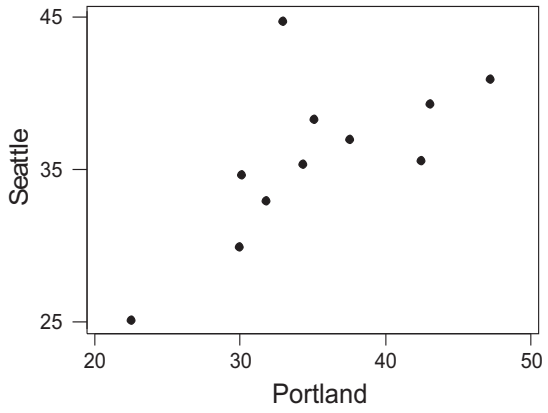
menyelidiki hubungan tersebut, berikut diberikan data jumlah curah hujan tahunan (dalam inci) pada Tabel 4.1.

Tabel 4.1. Data Curah Hujan

Tahun	Curah hujan di Seattle (Y)	Curah hujan di Portland (X)
1980	35.60	42.41
1981	35.40	34.29
1982	39.32	43.04
1983	40.93	47.19
1984	36.99	37.50
1985	25.13	22.48
1986	38.34	35.04
1987	29.93	29.91
1988	32.98	31.72
1989	34.69	30.05
1990	44.75	32.86

Gambar 4.1 memperlihatkan letak 11 titik data dengan sumbu x adalah curah hujan di Portland dan sumbu y adalah curah hujan di Seattle. Karena titik-titik tersebut membentuk suatu pola yaitu mendekati linear, maka model regresi linear $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ bisa digunakan untuk melihat seberapa baik curah hujan di Seattle dapat dijelaskan sebagai fungsi linear curah hujan di Portland.

Berdasar analisis regresi awal terhadap curah hujan tahunan di Seattle dan Portand untuk tahun 1950 sampai dengan 1979 diperoleh beberapa informasi mengenai nilai parameter. Informasi *prior* tersebut ditunjukkan dengan indeks a yaitu $(\hat{\beta}_0)_a = 5.513$, $(\hat{\beta}_1)_a = 0.8961$, $\text{SD}(\hat{\beta}_0)_a = 1.140$, $\text{SD}(\hat{\beta}_1)_a = 0.02986$ dan korelasi$((\hat{\beta}_0)_a, (\hat{\beta}_1)_a)$ = - 0.9871. Agar persamaan (4.5) bisa digunakan, maka dibentuk $\mu_a = (\beta_0)_a + 35.14(\beta_1)_a$ dengan 35.14 merupakan rata-rata curah hujan di Portland untuk tahun 1980 sampai 1990, sehingga $\hat{\mu}_a = 5.513 + (35.14)(0.8961) = 37.00$ dan $\text{SD}(\hat{\mu}_a) = 0.1977$. Nilai 0.1977 adalah hasil akar dari $(1.410)^2 + (35.14)^2(0.02986)^2 + 2(35.14)(1.140)(0.02986)(-0.9871)$. Distribusi



Gambar 4.1. Curah hujan di Seattle dan Portland

prior yang ditentukan adalah harga harapan μ, $e_\mu = 37.00$ dan deviasi standar μ, $c_\mu = 0.1977$ serta harga harapan β_1, $e_{\beta_1} = 0.8961$ dan deviasi standar β_1, $c_{\beta_1} = 0.02986$. Selanjutnya, dari data sekarang diperoleh $\bar{y} = 35.82$, $(\hat{\beta}_1)_{LS} = 0.5063$ dan $\Sigma(x_i - \bar{x})^2 = 497.2$. Berdasar persamaan (4.5) dihitung estimasi Bayes untuk β_0 dan β_1, diperoleh $\hat{\beta}_1 = 0.7764$, $\hat{\mu} = 36.65$ dan $\hat{\beta}_0 = 9.363$. Jadi, estimasi Bayes untuk garis regresi adalah $\hat{Y} = 9.363 + 0.7764X$.

Jika tidak terdapat informasi *prior* tentang nilai parameter, maka estimasi Bayes untuk β_0 dan β_1 sama dengan estimasi kuadrat terkecil. Berdasar persamaan (4.3) dan (4.4) serta *software* SPSS 11 diperoleh $\hat{\beta}_0 = 18.034$ dan $\hat{\beta}_1 = 0.506$, sehingga estimasi garis regresi adalah $\hat{Y} = 18.034 + 0.506X$.

Setelah diperoleh estimasi model regresi, selanjutnya akan diuji apakah terdapat hubungan yang signifikan antara curah hujan tahunan di Portland dan Seattle. Untuk mengetahui hal tersebut maka dilakukan uji hipotesis sebagai berikut.

1. $H_0 : \beta_1 = 0$ (tidak terdapat hubungan linear antara curah hujan tahunan di Seattle dan Portland)

   $H_1 : \beta_1 \neq 0$ (terdapat hubungan linear antara curah hujan tahunan di Seattle dan Portland).

2. Daerah kritis, H_0 ditolak jika $\text{Prob}(H_0|\boldsymbol{y}) < \text{Prob}(H_0) = \frac{1}{2}$.

3. Dengan *software SPSS 11*, diperoleh r = 0.669 sehingga g = 0.0611. Selanjutnya, dengan menggunakan persamaan (4.7) diperoleh $\text{Prob}(H_0|\boldsymbol{y}) =$ 0.1982.

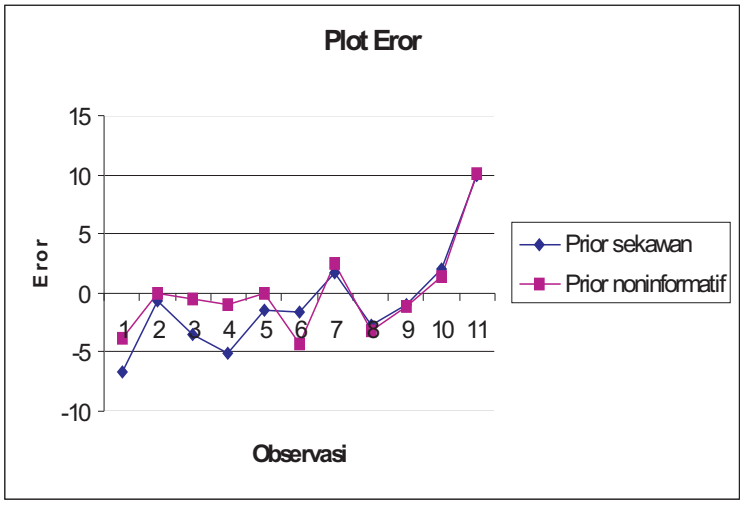
4. Karena $\text{Prob}(H_0|\boldsymbol{y})$= 0.1982 $< \text{Prob}(H_0) = \frac{1}{2}$ maka H_0 ditolak. Jadi, terdapat hubungan linear antara curah hujan tahunan di Seattle dan Portland.

Berdasar distribusi *prior* yang digunakan, diperoleh estimasi model regresi linear $\hat{Y}_{sekawan} = 9.363 + 0.7764X$ dan $\hat{Y}_{noninformatif} = 18.034 + 0.506X$. Selanjutnya, dilakukan analisis eror untuk membandingkan kedua model tersebut. Perhitungan eror, $y - \hat{y}$, berdasar distribusi *prior* sekawan dan distribusi *prior* noninformatif disajikan pada Tabel 4.2.

Tabel 4.2. Eror Regresi Linear Sederhana Berdasar Distribusi *Prior* Sekawan dan Distribusi *Prior* Noninformatif

y	$\hat{y}_{sekawan}$	$\hat{y}_{noninformatif}$	$e_{sekawan}$	$e_{noninformatif}$
35.60	42.290124	39.49346	-6.690124	-3.89346
35.40	35.985756	35.38474	-0.585756	0.01526
39.32	42.779256	39.81224	-3.459256	-0.49224
40.93	46.001316	41.91214	-5.071316	-0.98214
36.99	38.478	37.009	-1.488	-0.019
25.13	26.816472	29.40888	-1.686472	-4.27888
38.34	36.58056	35.76424	1.771944	2.57576
29.93	32.585124	33.16846	-2.655124	-3.23846
32.98	33.990408	34.08432	-1.010408	-1.10432
34.69	32.69382	33.2393	1.99618	1.4507
44.75	34.875504	34.66116	9.874496	10.08884

Gambar 4.2 memperlihatkan plot eror berdasar distribusi *prior* sekawan dan distribusi *prior* noninformatif. Dari Gambar 4.2 terlihat bahwa eror berdasar distribusi *prior* sekawan relatif lebih kecil daripada eror berdasar distribusi *prior*



Gambar 4.2. Plot eror regresi linear sederhana berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan

noninformatif. Pernyataan tersebut berarti bahwa terdapat perbedaan antara estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan. Untuk mengetahui apakah pernyataan tersebut benar atau salah maka harus dilakukan uji hipotesis. Uji hipotesis yang digunakan untuk membandingkan $\hat{Y}_{sekawan}$ dan $\hat{Y}_{noninformatif}$ adalah uji t untuk dua sampel berpasangan. Selanjutnya, akan diuji apakah mean eror berdasar distribusi *prior* sekawan lebih kecil daripada mean eror berdasar distribusi *prior* noninformatif. Uji hipotesis yang dilakukan adalah sebagai berikut.

1. $H_0 : \mu_{sekawan} = \mu_{noninformatif}$ (mean eror berdasar distribusi *prior* sekawan sama dengan mean eror berdasar distribusi *prior* noninformatif)
   $H_1 : \mu_{sekawan} < \mu_{noninformatif}$ (mean eror berdasar distribusi *prior* sekawan lebih kecil daripada mean eror berdasar distribusi *prior* noninformatif)

2. Dipilih tingkat signifikansi $\alpha = 0.05$.

3. Daerah kritis, H_0 ditolak jika $t < -t_{((n-1),\,\alpha)}$.
   Dari tabel t diperoleh $t_{((11-1),\,0.05)} = 1.812$.

4. Statistik uji.
   Berdasar *output software SPSS 11*, diperoleh $t = -1.443$

5. Karena -1.443 > - 1.812, maka H_0 tidak ditolak. Jadi mean eror berdasar distribusi *prior* sekawan sama dengan mean eror berdasar distribusi *prior* noninformatif. Dengan kata lain, tidak terdapat perbedaan antara estimasi model regresi berdasar distribusi *prior* sekawan dan estimasi model regresi berdasar distribusi *prior* noninformatif.

4.6.2 Contoh Kasus Regresi Linear Ganda

Suatu penelitian dilakukan untuk mengetahui pengaruh beberapa faktor terhadap bekas roda di jalan aspal (Birkes dan Dodge [3]). Faktor-faktor tersebut adalah

X_1 : kekentalan aspal yang ditransformasi dengan fungsi logaritma

X_2 : persentase aspal di permukaan jalan

X_3 : persentase aspal di dasar jalan

X_4 : persentase *fines* di permukaan jalan

X_5 : persentase *voids* di permukaan jalan

Variabel tak bebas Y merupakan logaritma perubahan kedalaman bekas roda (dalam inci) setiap satu juta roda yang melewati jalan aspal. Model regresinya adalah $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$. Data diberikan pada Tabel 4.3.

Berdasar informasi dari data awal, diperoleh vektor mean $\boldsymbol{b}$ dan matriks $\boldsymbol{V}$ sebagai berikut

$$\boldsymbol{b} = (-3.55,\ -0.44,\ 0.64,\ 0.13,\ 0.041,\ 0.14)$$

$$\boldsymbol{V} = \begin{pmatrix} 1690 & -38 & -174 & -110 & -10 & -47 \\ -38 & 4.07 & 3.46 & 3.23 & 0.47 & 0.23 \\ -174 & 3.46 & 19.81 & 8.88 & 1.15 & 5.31 \\ -110 & 3.23 & 8.88 & 12.47 & -0.17 & 1.56 \\ -10 & 0.47 & 1.15 & -0.17 & 0.60 & 0.15 \\ -47 & 0.23 & 5.31 & 1.56 & 0.15 & 2.69 \end{pmatrix}$$

Tabel 4.3. Data Bekas Roda di Jalan Aspal

No	Log bekas roda (Y)	X_1	X_2	X_3	X_4	X_5
1	-0.119	1.944	4.97	4.66	6.5	4.625
2	0.130	1.792	5.01	4.72	8.0	4.977
3	0.158	1.699	4.96	4.90	6.8	4.322
4	0.204	1.763	5.20	4.70	8.2	5.087
5	0.041	1.954	4.80	4.60	6.6	5.971
6	-0.071	1.820	4.98	4.69	6.4	4.647
7	0.079	2.146	5.35	4.76	7.3	5.115
8	-0.252	2.380	5.04	4.80	7.8	5.939
9	-0.143	2.623	4.80	4.80	7.4	5.916
10	-0.328	2.699	4.83	4.60	6.7	5.471
11	-0.481	2.255	4.66	4.72	7.2	4.602
12	-0.585	2.431	4.67	4.50	6.3	5.043
13	-0.119	2.230	4.72	4.70	6.8	5.075
14	-0.097	1.991	5.00	5.07	7.2	4.334
15	0.301	1.544	4.70	4.80	7.7	5.705

Sebelum memperoleh estimasi model regresi, terlebih dahulu dihitung estimator Bayes dengan menggunakan persamaan (4.13) dan *software Minitab 11*, diperoleh

$$\hat{ß}_{Bayes} = (-2.713,\ -0.5748,\ 0.4410,\ 0.1706,\ 0.00207,\ 0.1636).$$

Jadi, estimasi model regresinya adalah $\hat{Y} = -2.713 - 0.5748X_1 + 0.4410X_2 + 0.1706X_3 + 0.00207X_4 + 0.1636X_5$.

Jika tidak terdapat informasi *prior* tentang nilai parameter, maka estimator Bayes untuk $\hat{ß}$ sama dengan estimasi kuadrat terkecil. Berdasar persamaan (4.8) dan *software Minitab 11* diperoleh

$$\hat{ß}_{LS} = (-3.36,\ -0.582,\ 0.353,\ 0.383,\ -0.0091,\ 0.196).$$

Jadi, estimasi model regresinya adalah $\hat{Y} = -3.36 - 0.582X_1 + 0.353X_2 + 0.383X_3 - 0.0091X_4 + 0.196X_5$.

Setelah diperoleh estimasi model regresi, selanjutnya akan diuji apakah terdapat hubungan yang signifikan antara perubahan kedalaman bekas roda dan kelima faktor yang mempengaruhinya. Dari Tabel 4.3 diperoleh $n = 15$, $p = 5$ dan $q = 0$. Uji hipotesis yang dilakukan adalah sebagai berikut.

1. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (kekentalan aspal, persentase aspal di permukaan jalan, persentase aspal di dasar jalan, persentase *fines* di permukaan jalan dan persentase *voids* di permukaan jalan tidak berpengaruh terhadap perubahan kedalaman bekas roda)
   H_1 : paling tidak ada satu $\beta_i \neq 0$, $i = 1, 2, 3, 4, 5$ (paling tidak ada satu di antara kekentalan aspal, persentase aspal di permukaan jalan, persentase aspal di dasar jalan, persentase *fines* di permukaan jalan dan persentase *voids* di permukaan jalan yang berpengaruh terhadap perubahan kedalaman bekas roda)

2. Daerah kritis, H_0 ditolak jika $\text{Prob}(H_0|\boldsymbol{y}) < \text{Prob}(H_0) = \frac{1}{2}$

3. Dengan *software Minitab 11* dan persamaan (4.16) diperoleh $F_{LS} = 9.315$ dan $R^2 = 0.838$, sehingga $g = 0.0004$. Jadi, dengan persamaan (4.15) diperoleh $\text{Prob}(H_0|\boldsymbol{y}) = 0.019$.

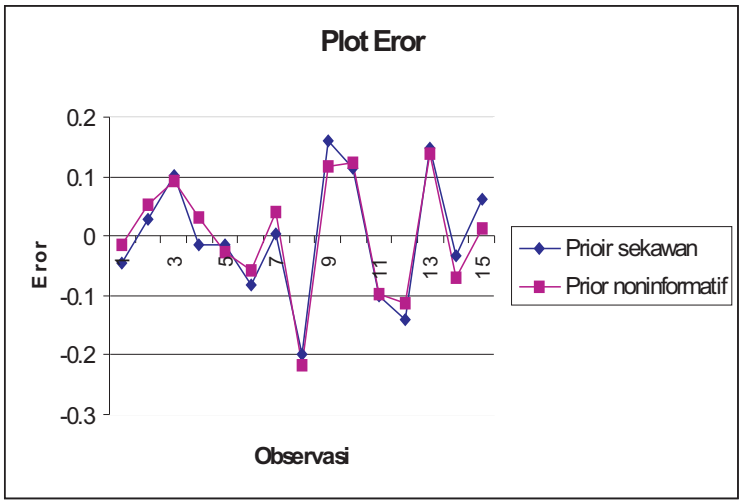
4. Karena $\text{Prob}(H_0|\boldsymbol{y}) = 0.019 < \text{Prob}(H_0) = \frac{1}{2}$ maka H_0 ditolak. Jadi, paling tidak ada satu di antara kekentalan aspal, persentase aspal di permukaan jalan, persentase aspal di dasar jalan, persentase *fines* di permukaan jalan dan persentase *voids* di permukaan jalan yang berpengaruh terhadap perubahan kedalaman bekas roda.

Berdasar distribusi *prior* yang digunakan, diperoleh estimasi model regresi linear $\hat{Y}_{sekawan} = -2.713 - 0.5748X_1 + 0.4410X_2 + 0.1706X_3 + 0.00207X_4 + 0.1636X_5$ dan $\hat{Y}_{noninformatif} = -3.36 - 0.582X_1 + 0.353X_2 + 0.383X_3 - 0.0091X_4 + 0.196X_5$. Selanjutnya, dilakukan analisis eror untuk membandingkan kedua model tersebut. Perhitungan eror, $y - \hat{y}$, berdasar distribusi *prior* sekawan dan distribusi *prior* noninformatif disajikan pada Tabel 4.4.

Tabel 4.4. Eror Regresi Linear Ganda Berdasar Distribusi *Prior* Sekawan dan Distribusi *Prior* Noninformatif

y	$\hat{y}_{sekawan}$	$\hat{y}_{noninformatif}$	$e_{sekawan}$	$e_{noninformatif}$
-0.119	-0.0735402	-0.104868	-0.0454598	-0.014132
0.130	0.1023976	0.076038	0.0276024	0.053962
0.158	0.05487	0.063994	0.10313	0.094006
0.204	0.2178548	0.172066	-0.0138548	0.031934
0.041	0.0559184	0.069228	-0.0149184	-0.028228
-0.071	0.0106552	-0.012458	-0.0816552	-0.05542
0.079	0.0768102	0.038768	0.0021898	0.040232
-0.252	-0.0517376	-0.034576	-0.2002624	-0.217424
-0.143	-0.3018448	-0.26159	0.1588448	0.11859
-0.328	-0.4406706	-0.452682	0.1126076	0.124682
-0.481	-0.3810908	-0.383198	-0.0999092	-0.097802
-0.585	-0.445093	-0.471734	-0.139907	-0.113266
-0.119	-0.267118	-0.25878	0.148118	0.13978
-0.097	-0.0635384	-0.028008	-0.0334616	-0.068992
0.301	0.2403658	0.287002	0.0606342	0.013998

Gambar 4.3 memperlihatkan plot eror berdasar distribusi *prior* sekawan dan distribusi *prior* noninformatif. Dari Gambar 4.2 terlihat bahwa eror berdasar distribusi *prior* sekawan relatif lebih kecil daripada eror berdasar distribusi *prior* noninformatif. Pernyataan tersebut berarti bahwa terdapat perbedaan antara estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan. Untuk mengetahui apakah pernyataan tersebut benar atau salah maka harus dilakukan uji hipotesis. Uji hipotesis yang digunakan untuk membandingkan $\hat{Y}_{sekawan}$ dan $\hat{Y}_{noninformatif}$ adalah uji t untuk dua sampel berpasangan. Selanjutnya, akan diuji apakah mean eror berdasar distribusi *prior* sekawan lebih kecil daripada mean eror berdasar distribusi *prior* noninformatif. Uji hipotesis yang dilakukan adalah sebagai berikut.



Gambar 4.3. Plot eror regresi linear ganda berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan

1. $H_0 : \mu_{sekawan} = \mu_{noninformatif}$ (mean eror berdasar distribusi *prior* sekawan sama dengan mean eror berdasar distribusi *prior* noninformatif)
   $H_1 : \mu_{sekawan} < \mu_{noninformatif}$ (mean eror berdasar distribusi *prior* sekawan lebih kecil daripada mean eror berdasar distribusi *prior* noninformatif).

2. Dipilih tingkat signifikansi $\alpha = 0.05$

3. Daerah kritis, H_0 ditolak jika $t < -t_{((n-1),\,\alpha)}$.
   Dari tabel t diperoleh $t_{((15-1),\,0.05)} = 1.761$.

4. Berdasar *output software SPSS 11*, diperoleh $t = -0.332$

5. Karena -0.332 > - 1.761, maka H_0 tidak ditolak. Jadi mean eror berdasar distribusi *prior* sekawan sama dengan mean eror berdasar distribusi *prior* noninformatif. Dengan kata lain, tidak terdapat perbedaan antara estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasar pembahasan, diperoleh kesimpulan sebagai berikut.

1. Pada kasus regresi linear sederhana, jika informasi *prior* tentang parameter tidak diketahui, maka estimasi Bayes sama dengan estimasi kuadrat terkecil yaitu

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \ , \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{(\sum_{i=1}^{n} x_i) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \ .$$

Jika informasi *prior* tentang parameter diketahui, maka estimasi Bayes adalah

$$\hat{\beta}_0 = \hat{\mu} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \{\frac{c_{\beta_1}^{-2}}{c_{\beta_1}^{-2} + \sum(x_i - \bar{x}^2)}\} e_{\beta_1} + \{\frac{\sum(x_i - \bar{x}^2)}{c_{\beta_1}^{-2} + \sum(x_i - \bar{x}^2)}\} (\hat{\beta}_1)_{LS}$$

dengan

$$\hat{\mu} = \{\frac{c_{\mu}^{-2}}{c_{\mu}^{-2} + n}\} e_{\mu} + \{\frac{n}{c_{\mu}^{-2} + n}\} \bar{y}$$

dan $(\hat{\beta}_1)_{LS}$ merupakan estimasi kuadrat terkecil untuk data sekarang.

Pada kasus regresi linear ganda, jika informasi *prior* tentang parameter tidak diketahui, maka estimasi Bayes sama dengan estimasi kuadrat terkecil yaitu $\hat{ß} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, jika informasi *prior* tentang parameter diketahui, maka estimasi Bayes adalah $\hat{ß}_{Bayes} = \boldsymbol{V}_* \boldsymbol{V}^{-1} \boldsymbol{b} + \boldsymbol{V}_* \boldsymbol{X}' \boldsymbol{X} \hat{ß}_{LS}$.

2. Uji signifikansi parameter regresi linear sederhana dan regresi linear ganda dengan metode Bayesian dilakukan dengan menghitung probabilitas *posterior* hipotesis nol, yaitu

$$Prob(H_0|\boldsymbol{y}) = \frac{1}{1 + \frac{1}{\sqrt{g}}} \ .$$

Jika nilai $Prob(H_0|\boldsymbol{y}) < Prob(H_0) = \frac{1}{2}$ maka H_0 ditolak.

3. Berdasar analisis eror yang dilakukan pada suatu kasus disimpulkan bahwa tidak terdapat perbedaan antara estimasi model regresi berdasar distribusi *prior* noninformatif dan distribusi *prior* sekawan.

5.2 Saran

Bagian terpenting dari analisis data dengan metode Bayesian adalah pemilihan distribusi *prior*. Setelah melakukan estimasi parameter regresi dengan metode Bayesian, terlihat bahwa hasil akhir sangat bergantung pada pemilihan distribusi *prior*. Bagi pembaca yang berminat mengestimasi parameter regresi dengan metode Bayesian bisa memilih distribusi *prior* yang berbeda.

DAFTAR PUSTAKA

[1] Anton, H. *Elementary Linear Algebra*, fifth ed., John Wiley and Sons, Inc., New York, 1992.

[2] Bain, L. J. and M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, second ed., Duxbury Press, Inc., California, 1992.

[3] Birkes, D. and Y. Dodge, *Alternative Method of Regression*, John Wiley and Sons, Inc., New York, 1993.

[4] Johnson, R. A. and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice - Hall, Inc., New Jersey, 1982.

[5] Johnson, R. A. and G. K. Bhattacharyya, *Statistics Principles and Methods*, third ed., John Wiley and Sons, Inc., New York, 1996.

[6] Larson, H. J. *Introduction to Probability Theory and Statistical Inference*, John Wiley and Sons, Inc., New York, 1974.

[7] Montgomery, D. C. and E. A. Peck, *Introduction to Linear Regression Analysis*, second ed., John Wiley and Sons, Inc., New York, 1992.

[8] Seber, G. *Linear Regression Analysis*, John Wiley and Sons, Inc., New York, 1977.

[9] Sembiring, R. K. *Analisis Regresi*, ITB, Bandung, 1995.

[10] Soejoeti, Z. dan Subanar, *Inferensi Bayesian*, Karunika, Universitas Terbuka, Jakarta, 1988.

[11] Walpole, R. E. dan R. H. Myers, *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan,* Terjemahan R.K Sembiring, ITB, Bandung, edisi kedua, ITB, Bandung, 1995.

LAMPIRAN

Lampiran 1 : *Ouput SPSS 11* untuk data curah hujan di Seattle dan Portland.

Lampiran 2 : Uji t untuk dua sampel berpasangan pada data curah hujan di Seattle dan Portland.

Lampiran 3 : *Output Minitab 11* untuk data bekas roda di jalan aspal.

Lampiran 4 : Uji t untuk dua sampel berpasangan pada data bekas roda di jalan aspal.

Lampiran 5 : Artikel *"Bayesian Method On Linear Regression Model"*.

Lampiran 1 : Bukti Teorema 2.1.1, Teorema 2.1.2 dan Teorema 2.1.3.

Teorema 2.1.1. *Jika A dan B adalah dua kejadian sembarang, maka*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Bukti. Kejadian $A \cup B$ dan kejadian A masing-masing merupakan gabungan dari kejadian-kejadian yang saling asing. Berdasar sifat himpunan dapat ditunjukkan bahwa

$$A \cup B = (A \cap B^{'}) \cup B \quad dan \quad A = (A \cap B) \cup (A \cup B^{'})$$

Kejadian $A \cap B^{'}$ dan B adalah dua kejadian saling asing karena $(A \cap B^{'}) \cap B = \emptyset$, sehingga $P(A \cup B) = P(A \cap B^{'}) + P(B)$. Demikian juga, kejadian $A \cap B$ dan B adalah dua kejadian saling asing, sehingga $P(A) = P(A \cap B) + P(B)$. Jadi

$$\begin{aligned} P(A \cup B) &= P(A \cap B^{'}) + P(B) \\ &= [P(A) - P(A \cap B)] + P(B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

□

Teorema 2.1.2. *Misal kejadian $B_1, B_2, \ldots, B_k$ merupakan kejadian yang saling asing dari ruang sampel S dengan $P(B_i) \neq 0$ untuk $i = 1, 2, \ldots, k$, maka untuk setiap kejadian A anggota S*

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

Bukti. Kejadian A merupakan gabungan dari sejumlah kejadian $B_1 \cap A, B_2 \cap A, \ldots, B_k \cap A$ yang saling asing, sehingga

$$\begin{aligned} P(A) &= P[(B_1 \cap A) \cup (B_2 \cap A) \cup \ldots \cup (B_k \cap A)] \\ &= P((B_1 \cap A)) + P((B_2 \cap A)) + \cdots + P((B_k \cap A)) \\ &= \sum_{i=1}^{k} P(B_i \cap A) \\ &= \sum_{i=1}^{k} P(B_i)P(A|B_i) \end{aligned}$$

□

Teorema 2.1.3. *Misal kejadian $B_1, B_2, \ldots, B_k$ merupakan kejadian yang saling asing dari ruang sampel S dengan $P(B_i) \neq 0$ untuk $i = 1, 2, \ldots, k$. Misalkan A suatu kejadian sembarang dalam S dengan $P(A) \neq 0$, maka*

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

untuk $r = 1, 2, \ldots, k$.

Bukti. Berdasar definisi peluang bersyarat, maka

$$P(B_r|A) = \frac{P(B_r \cap A)}{P(A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}\,.$$

□

Lampiran 2 : *Output SPSS 11* untuk data curah hujan di Sattle dan Portland.

Lampiran 3 : Uji t untuk dua sampel berpasangan

Lampiran 4 : *Output Minitab 11* untuk data bekas roda di jalan aspal.

Notasi matriks di bawah ini digunakan untuk memperoleh $\hat{ß}_{Bayes}$.

M1 : matriks $\boldsymbol{X}$

M2 : tranpose matriks $\boldsymbol{X}$

M3 : matriks $\boldsymbol{X}'\boldsymbol{X}$

M4 : matriks $\boldsymbol{V}$

M5 : matriks $\boldsymbol{V}^{-1}$

M6 : jumlah M5 dan M3

M7 : matriks $\boldsymbol{V}_*$

M8 : vektor mean $\boldsymbol{b}$

M9 : vektor $\hat{ß}_{LS}$

M10: perkalian M7 dan M4

M11: perkalian M10 dan M8

M12: perkalian M7 dan M3

M13: perkalian M12 dan M9

M14: jumlah M11 dan M13

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.669[a]	.448	.387	4.17539

a. Predictors: (Constant), X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	127.457	1	127.457	7.311	.024[a]
	Residual	156.905	9	17.434		
	Total	284.362	10			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.034	6.699		2.692	.025
	X	.506	.187	.669	2.704	.024

a. Dependent Variable: Y

```
MTB > Read 15 6 M1
DATA> 1 1.944 4.97  4.66  6.5   4.625
DATA> 1 1.792 5.01  4.72  8.0   4.977
DATA> 1 1.699 4.96  4.90  6.8   4.322
DATA>1 1.763 5.20  4.70  8.2   5.087
DATA> 1 1.954 4.80  4.60  6.6   5.971
DATA> 1 1.820 4.98  4.69  6.4   4.647
DATA> 1 2.146 5.35  4.76  7.3   5.115
DATA> 1 2.380 5.04  4.80  7.8   5.939
DATA> 1 2.623 4.80  4.80  7.4   5.916
DATA> 1 2.699 4.83  4.60  6.7   5.471
DATA> 1 2.255 4.66  4.72  7.2   4.602
DATA> 1 2.431 4.67  4.50  6.3   5.043
DATA> 1 2.230 4.72  4.70  6.8   5.075
DATA> 1 1.991 5.00  5.07  7.2   4.334
DATA> 1 1.544 4.70  4.80  7.7   5.705
    15 rows read.
MTB > Transpose M1 M2.
MTB > Multiply M2 M1 M3.

MTB > Read 6 6 M4.
DATA> 1690 -38 -174 -110 -10 -47
DATA> -38 4.07 3.46 3.23 0.47 0.23
DATA> -174 3.46 19.81 8.88 1.15 5.31
DATA> -110 3.23 8.88 12.47 -0.17 1.56
DATA> -10 0.47 1.15 -0.17 0.60 .15
DATA> -47 0.23 5.31 1.56 0.15 2.69
    6 rows read.
MTB > Invert M4 M5.
MTB > Add M5 M3 M6.
MTB > Invert M6 M7.
MTB > Read 6 1 M8.
DATA> -3.55
DATA> -0.44
DATA> 0.64
DATA> 0.13
DATA> 0.041
DATA> 0.14
    6 rows read.
MTB > Read 6 1 M9.
DATA> -3.362
DATA> -0.5817
DATA> 0.3529
DATA> 0.3831
DATA> -0.009064
DATA> 0.1964
    6 rows read.
```

```
MTB > Multiply M7 M5 M10.
MTB > Multiply M10 M8 M11.
MTB > Multiply M7 M3 M12.
MTB > Multiply M12 M9 M13.
MTB > Add M11 M13 M14.
MTB > prin M14
```

Data Display

```
Matrix M14

-2.71271
-0.57480
 0.44094
 0.17064
 0.00207
 0.16357
```